DOCUMENT RESUME

ED 334 272 TM 016 920

AUTHOR Foong, Yoke-Yeen; Lam, Tit-Loong
TITLE The Use of the Graded Response Model in Computerized Adaptive Testing of the Attitudes to Science Scale.
PUB DATE Apr 91
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; *Attitude Measures; *Computer Assisted Testing; Computer Simulation; Foreign Countries; Grade 8; Junior High Schools; *Junior High School Students; Mathematical Models
IDENTIFIERS Attitudes To Science Scale; *Graded Response Model; Singapore; *Two Stage Testing

ABSTRACT

The graded response model for two-stage testing was applied to an attitudes toward science scale using real-data simulation. The 48-item scale was administered to 920 students at a grade-8 equivalent in Singapore. A two-stage 16-item computerized adaptive test was developed. In two-stage testing an initial, or routing, test is followed by a second-stage testlet of greater or lesser difficulty based on performance. A conventional test of the same length as the adaptive two-stage test was selected from the 48-item pool. Responses to the conventional test, the routing test, and a testlet were simulated. The algorithm of E. Balas (1965) and the multidimensional knapsack problem of optimization theory were used in test development. The simulation showed the efficiency and accuracy of the two-stage test with the graded response model in estimating attitude trait levels, as evidenced by better results from the two-stage test than its conventional counterpart and the reduction to one-third of the length of the original measure. Six tables and three graphs are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334272

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

YOKE-YEEN FOONG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE USE OF THE GRADED RESPONSE MODEL IN COMPUTERIZED ADAPTIVE TESTING OF THE ATTITUDES TO SCIENCE SCALE

Yoke-Yeen Foong*
Tit-Loong Lam

University of Massachusetts at Amherst

Paper presented at the 1991 Annual Meeting of the American Educational Research Association, Chicago

*Both authors on study leave from the National Institute of Education, Singapore.

BEST COPY AVAILABLE

The Use of the Graded Response Model in Computerized Adaptive Testing of the Attitudes to Science Scale

Attitude scales often make use of Likert formats which are scored into ordered categories represented by integers along the negative - positive attitude continuum. The precision of attitude trait estimates will be maximized by making use of the information contained in each item response. This study attempts to extend previous work by applying Samejima's (1969) graded response model for two-stage testing to attitudes to science scale using real-data simulation. The algorithm of Balas and multidimensional knapsack problem of Optimization Theory will be used in the development of tests. The study shows the efficiency and accuracy in using well designed two-stage test with the graded response model in estimation of attitude trait levels. This is indicated by the better results of the two-stage test compared to its conventional counterpart, and a reduction of the test length to one third its original length.

THE USE OF THE GRADED RESPONSE MODEL IN COMPUTERIZED ADAPTIVE TESTING OF THE ATTITUDES TO SCIENCE SCALE

Yoke-Yeen Foong and Tit-Loong Lam
University of Massachusetts at Amherst

Attitude scales often make use of Likert formats which are scored into ordered categories represented by integers along the negative - positive attitude continuum. One major limitation of computerized adaptive testing (CAT) is the reliance on dichotomous item response data. The binary scoring of items results in a loss of information that might otherwise be obtained from polychotomous scoring. The precision of attitude trait estimates can be maximized by using the information contained in each item response. Several item response models for polychotomous data are available, including Samejima's (1969) graded response model.

The Problem

Much work has been done using simulated data for CAT with binary scoring of cognitive data. Adema (1990) and Theunissen (1985, 1986) reported studies on two-stage testing for simulated dichotomous data. Some basic research has also been conducted on CAT procedures using item response models for polychotomous data with simulated data.

This study attempts to extend previous work by applying the graded response model for two-stage testing to attitudes to science scale using real-data simulation. The use of life data in the simulation study will give some measure of confidence in generalizing the results of the study in the real world and

provide further insights to the design of CAT software which uses the polychotomous model.

Information

Building upon Birnbaum's (1968) formulation of information functions, Samejima defined item information for the case where items are scored polychotomously as

$$I_i(\theta) = \Sigma \frac{P'_{x_i}(\theta)^2}{P_{x_i}(\theta)}.$$

$P_{x_i}(\theta)$ is the probability of receiving category score xi conditional on θ, and $P'_{x_i}(\theta)$ is the derivative of $P_{x_i}(\theta)$. The information an item has depends on the item discrimination parameter and the category boundary parameters. If the category boundary parameters span a wide range, the item provides more information across a wide range of the θ scale than items with category boundaries that span a small range. Items with high discrimination give more precise measurement of trait levels in the range of the category boundary parameters than items with low discrimination. The information provided by a test at θ is simply the sum of the item information functions at θ, and the items contribute independently to the test information function.

Two-stage testing

Work in the field of adaptive testing in the 1970s and 1980s has been in the testing of algorithms that adaptively select each item in the test, one item at a time (Thissen & Mislevy, 1990). Thissen and Mislevy (1990) proposed that measurement can be as good with testing algorithms that adapt less often than every

time an item is presented, and that such systems known as two-stage testing algorithms may be much more practical.

In the 'first-stage test' or 'routing test', the examinee is first administered a single fixed test which is then scored. Following this, another test known as the 'second-stage test' is administered. Those who perform better in the routing test will be administered a more difficult second-stage test or testlet, while those who perform less well will be administered an easier second-stage testlet. This is an adaptive testing algorithm that adapts exactly once, between the several-item-long routing test and the often longer second-stage testlet.

Method

Instrument

The attitudes to science scale consisting of 48 items was designed to measure students' interest, enjoyment, and motivation in science, their career preference, their perception of the difficulty in science, and their attitudes towards the social implications of science. Each item in the instrument was rated on a five-point Likert scale.

The instrument was administered to 920 secondary two students (equivalent to Grade 8) from various social, cultural and educational background, representative of that population in Singapore.

Since one of the basic assumptions of the graded response model is that the underlying trait being measured by the set of

items is unidimensional, the item responses were subjected to principal components and maximum likelihood factor analyses to investigate its dimensionality.

Parameter Estimation

The category boundary and discrimination parameters were estimated for the 48 items in the attitudes to science scale from MULTILOG (Thissen, 1988) using the marginal maximum likelihood method (Bock & Aitkin, 1981). Subsequently, theta values for the students were obtained from MULTILOG using maximum likelihood estimation with a Bayes modal solution. The item parameter estimates were used to compute the information for each item and thereby to construct the information tables.

CAT Procedures

In order to minimize misrouting to an inappropriate second-stage test, the target information function was specified at 4.8 (Betz and Weiss, 1973, 1974). This criterion for the routing test corresponds to a standard error which should not exceed 0.46. The classical reliability can be expressed as the intra class correlation, ρ and the standard error of measurement is given by

$$\sigma_\theta = \sigma_x \sqrt{1 - \rho}.$$

With the assumption that the error variance, $\theta = 1$, and the lower bound of $I(\theta) = 4.8$, $\sigma \approx 0.46$, and $\sigma^2 \approx 0.21$. This corresponds to a traditional reliability of 0.8.

Based on the work on the development of two-stage testing by Adema (1990) and Theunissen (1985, 1986), the item information was summed at five trait levels for the routing test. The test

design took the form of a multidimensional knapsack problem, one of the popular concepts in Optimization Theory as follows:

Minimize the objective function:

$$I_t = x_1 + x_2 + x_3 + \ldots\ldots + x_{n1}$$

subject to the constraints that:

$$I(\theta_j,1)x_1 + I(\theta_j,2)x_2 + I(\theta_j,3)x_3 \ldots\ldots + I(\theta_j,n1)x_{n1} \geq \text{Info}$$

where $x_i = 0$ or 1, $i = 1,2,\ldots n_1$, $j = 1,2,\ldots 5$, and n_1 is the size of the item bank.

The test information for the five specified attitude levels was denoted as total information, and the items were sorted according to the total information. The enumeration of the items involved optimal selection of the minimum number of items that would match the target information. Based on the initial attitude trait estimates from the routing test, the students were placed in 12 trait groups with the range of -2.5 to 3.0 with an interval of 0.5.

For the second-stage test, the design was such that given a certain test length, the items were chosen to give maximum information at the specified attitude segments.

Therefore, maximize:

$$I(\hat{\theta}_j,1)x_1 + I(\hat{\theta}_j,2)x_2 + \ldots\ldots + I(\hat{\theta}_j,j)x_{n2}$$

subject to the constraints:

$$x_1 + x_2 + x_3 + \ldots\ldots\ldots + x_{n2} = x_{nt}$$

where $x_i = 0$ or 1, $i = 1,2,3,\ldots,n_2$, $j = 1,2,3,\ldots,12$, n_2 is the size of the item bank stratum for the second-stage testing, and n_t is the length of the desired second-stage testlet.

There are 12 clusters of items or testlets in the second-stage testing to measure 12 equally spaced attitude segments to cover the whole attitude spectrum. Each testlet has maximum information at the midpoint of its ability range.

For this test design, the algorithm proposed by Balas (1965) was used. It involves a combinatorial approach to solving linear programs with zero-one variables. In fact, his method is a tree-search algorithm which uses information generated in the search to exclude portions of the tree under consideration. In computing Balas' algorithm and the enumerations, the software, STATA (Computing Resource Center, 1990) was used. The resulting response strings were generated to serve as data inputs for scoring using MULTILOG (Thissen, 1988).

A conventional test was developed with a test length equal to the sum of the length of the routing test and that of a testlet. The items were selected randomly from the same item pool. This conventional test was used as a basis for evaluating the effectiveness of the adaptive tests.

Analyses

Descriptive statistics were computed for the 48-item pool, 16-item conventional test, routing test, and the testlets.

The indices used in assessing the CAT strategies were:

1. the inaccuracy computed as:

$$IA(\theta) = \frac{\Sigma|\theta_h - \theta_e|}{N}$$

where θ_h, θ_e, and N are the estimated thetas (i.e. from

the two-stage testing), empirical thetas (i.e. from the conventional attitude instrument administration) and the number of respondents in the attitude groupings respectively.

2. the root mean square difference between the estimated and empirical thetas, given by:

$$RMSD(\theta) = \sqrt{\Sigma \frac{(\theta_e - \theta_o)^2}{N}}$$

Smaller RMSD or inaccuracy will imply estimates that are closer to the empirical thetas and thus, a greater level of accuracy in attitude estimation for that trait group.

3. the correlation analyses which show the degree of linear relationship between the estimated and the empirical thetas. De Ayala (1989) refers to the correlation coefficient as the 'validity coefficient' while Weiss and McBride (1984) refer to it as the fidelity coefficient.

Results

The descriptive statistics of the responses of the attitudes to science scale is shown in Table 1. An examination of the results shows that the items were of appropriate difficulty and the standard deviations of acceptable values.

Insert Table 1 about here

The results of principal components analysis of the responses to the attitudes to science scale showed that there was a single dominant factor. The eigenvalue for the first factor was 14.19 compared to a value of 2.37 for the second factor. The results of maximum likelihood factor analysis of the data showed that there was a single dominant factor which accounted for 72.35% of the common variance. These findings indicated that the attitudes to science scale met the unidimensional assumption at least approximately.

Parameter Estimation

Table 2 shows the graded response model estimated item parameters for the 48-item pool. Generally, the item discriminations were high although there were a few items with low discriminations. The category boundaries spanned a wide range.

Insert Table 2 and Figure 1 about here

Item Pool Information

The test information function for the item pool is shown in Figure 1. A multimodal information function was obtained. The test information was generally better at the lower theta values than at the higher levels. This function is better than a peaked information function for effective and efficient implementation of the CAT procedures. However, the maxima and minima of the test information function indicate that measurements were not of equal precision at all levels of the attitude trait levels. The best

test information function is one in which the function is very high and flat across a wide range of θ values.

The target information was set at 4.8 for the routing test. The solution of the knapsack problem resulted in a minimum test length of 6 items. The length of the testlets was set at 10 items. The graded response descriptive statistics of item parameters of the 48-item pool, 16-item conventional test, 6-item routing test, and 11 testlets are shown in Tables 3 and 4. The item discrimination and difficulties of the 48-item pool and the conventional pool showed a greater spread than the routing test and the testlets - a consequence of proper test design and enumeration technique. In addition, extreme b values were not selected for the routing test and the testlets. It can also be seen that the items for the 16-item conventional test covered a broad range of b values and thus, the associated attitude trait levels since no restrictions were imposed on its test design.

Insert Tables 3 and 4 about here

Table 5 and Figure 2 show a comparison of the inaccuracies of the 16-item conventional test and the two-stage test. The two-stage test showed consistently less inaccuracies than the same length conventional test for the whole range of abilities. In fact, except for the two extreme groups, the inaccuracies for the two-stage test were lower than the smallest inaccuracy for the conventional test. The smaller inaccuracies were a result of

correct routing of the respondents to their respective testlets.

The pattern of inaccuracies for the conventional test and the two-stage test was similar, except for the extreme ends. There was greater inaccuracies at the extreme ends for the conventional test, an indication that a greater measurement precision was found at the restricted range, but lesser precision at extreme range.

Insert Tables 5 & 6 and Figures 2 & 3 about here

Table 6 and Figure 3 show the root mean square differences of the attitude trait estimates for the 16-item conventional and the two-stage test. The results showed that the RMSD of the attitude trait estimates for the two-stage test were lower than the conventional test. The two tests had the same pattern for the RMSD for the attitude trait range.

The fidelity coefficient was 0.98 for the two-stage test compared to 0.95 for the 16-item conventional test.

Discussion

In general, the use of the graded response model theory in two-stage testing performed reasonably well. The model appeared to fit the item response data. Parameter estimates for all items and θ estimates for all respondents were obtained without any nonconvergence problems.

The use of graded response model theory in two-stage testing increased the precision in estimation of the attitude trait of respondents as shown by the efficiency of well designed two-stage test compared to the conventional test. The study showed that the respondents were accurately routed to the appropriate testlets even though the target information was set at a level of 4.8.

With the increased precision in estimation of the attitude trait and a high fidelity coefficient, this study shows that the use of the 16-item two-stage test would allow the gathering of just as much information as the 48-item test. The reduction of the test length of the conventional test to one third its original length will greatly reduce response burden. This compares well with typical CAT procedures which require less than half as many items to be administered to achieve the same level of measurement accuracy as the full length paper-and-pencil version. The results from this study support the findings of Dodd, Koch and De Ayala (1989) that substantially smaller item pools can be used successfully for polychotomous CAT compared to those required for dichotomous CAT. According to Dodd, et. al. (1989), item pools with as few as 30 items may be adequate for CAT, although item pools with a uniform distribution of information will perform better than peaked-information item pools.

The evaluative indices of this study - the small inaccuracies and the root mean square differences are comparable with the results obtained by previous studies on adaptive testing

using either the minimum standard error or the maximum information item selection (De Ayala, Dodd & Koch, 1990; Weiss, 1982).

Conclusions

The study shows the efficiency and accuracy in using well designed two-stage test with the graded response model in estimation of attitude trait levels. This is indicated by the better results of the two-stage test compared to an equal length conventional test, and a reduction of the test length to one third its original length. The most important practical finding of this study is that an item pool of 48 items can be used successfully for polychotomous CAT.

Bibliography

Adema, J.J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27,* 241-253.

Balas, E. (1965). An additive algorithm for solving linear programs with zero-one variables. *Operations Research, 13,* 517-546.

Betz, N.E. & Weiss, D.J. (1973). *An Empirical Study of Computerized Two-Stage Ability Testing* (Research Report 73-4). Minneapolis: University of Minnesota, Psychometrics Methods Program.

Betz, N.E. & Weiss, D.J. (1974). *Simulation Studies of Two-Stage Ability Testing* (Research Report 74-4). Minneapolis: University of Minnesota, Psychometrics Methods Program.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical Theories of Mental Test Scores.* Reading MA: Addison-Wesley.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika, 46,* 443-459.

Computing Resource Center (1990). *STATA Reference Manual*. Santa Monica, CA: Author.

De Ayala, R.J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement, 49,* 789-805.

De Ayala, R.J., Dodd, B.G. & Koch, W.R. (1990). A simulation and comparison of flexilevel and Bayesian computerized adaptive testing. *Journal of Educational Measurement, 27,* 227-239.

Dodd, B.G., Koch, W.R. & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129-143.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika, 50,* 411-420.

Theunissen, T.J.J.M. (1986). Optimization algorithms in test design. *Applied Psychological Measurement, 10,* 381-389.

Thissen, D. (1988). *MULTILOG Version 5.1 User's Guide.* Mooresville IN: Scientific Software, Inc.

Thissen, D.M. & Mislevy, R.J. (1990). Testing algorithms in H.Wainer, *Computerized Testing: A Primer.* New Jersey: Lawrence Erlbaum.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Weiss, D.J. & McBride, J.R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.

Table 1
Descriptive Statistics of the Responses of the Attitudes to Science Scale

Item	Mean	Std. Dev.	Min	Max
1	3.48	0.97	1	5
2	4.52	0.74	1	5
3	2.46	0.96	1	5
4	3.80	1.00	1	5
5	3.04	1.10	1	5
6	3.26	1.07	1	5
7	2.81	1.14	1	5
8	2.19	1.00	1	5
9	2.86	0.97	1	5
10	3.42	0.94	1	5
11	3.56	0.94	1	5
12	4.12	0.90	1	5
13	3.42	1.00	1	5
14	3.08	0.82	1	5
15	2.32	0.99	1	5
16	2.65	1.08	1	5
17	3.71	0.91	1	5
18	2.20	1.03	1	5
19	3.40	0.97	1	5
20	3.38	1.03	1	5
21	1.94	0.99	1	5
22	2.50	0.91	1	5
23	3.21	0.97	1	5
24	3.73	0.87	1	5
25	3.38	0.95	1	5
26	3.95	1.00	1	5
27	2.30	1.01	1	5
28	4.11	0.91	1	5
29	3.37	0.79	1	5
30	1.99	0.94	1	5
31	3.34	0.93	1	5
32	3.47	0.93	1	5
33	2.11	0.88	1	5
34	2.94	0.97	1	5
35	3.56	0.98	1	5
36	2.23	0.98	1	5
37	3.52	0.93	1	5
38	2.50	0.90	1	5
39	3.01	0.97	1	5
40	2.49	0.97	1	5
41	3.87	0.94	1	5
42	3.02	1.05	1	5
43	3.30	1.03	1	5
44	2.44	0.97	1	5
45	2.69	0.96	1	5
46	3.59	0.91	1	5
47	2.74	0.94	1	5
48	3.41	1.07	1	5

Table 2
Graded Response Model Estimated Item Parameters for the 48-Item Pool

Item	a	b_1	b_2	b_3	b_4
1	2.13	-2.35	-1.11	-0.06	1.85
2	0.80	-6.01	-4.92	-3.29	-0.55
3	1.03	-1.85	0.47	2.17	4.67
4	0.63	-5.93	-3.47	-1.10	1.96
5	1.50	-1.95	-0.74	1.07	2.02
6	0.30	-8.73	-3.44	0.44	7.72
7	1.60	-1.31	-0.43	1.25	2.28
8	0.75	-4.92	-2.95	-0.98	1.77
9	0.87	-2.90	-0.73	1.48	4.71
10	2.06	-2.35	-1.18	0.10	2.02
11	0.77	-4.94	-2.33	-0.57	3.23
12	0.90	-5.02	-3.65	-1.48	0.75
13	1.62	-2.45	-1.46	0.41	1.71
14	1.25	-2.90	-1.40	1.31	3.34
15	0.53	-6.99	-3.24	-1.31	3.16
16	1.45	-1.24	-0.27	1.74	2.73
17	1.87	-2.62	-1.74	-0.38	1.51
18	0.82	-4.10	-2.75	-0.97	1.67
19	1.60	-2.56	-1.34	0.22	2.00
20	1.14	-2.94	-1.17	-0.11	2.70
21	0.47	-7.53	-5.89	-2.08	0.91
22	0.86	-2.44	0.29	2.78	5.10
23	1.68	-2.39	-0.99	0.59	2.31
24	0.89	-5.28	-2.96	-0.61	2.28
25	1.64	-2.52	-1.43	0.38	2.00
26	0.46	-7.80	-4.71	-2.09	1.78
27	2.35	-2.43	-1.35	-0.30	1.16
28	0.61	-7.27	-5.01	-2.06	0.93
29	1.18	-4.22	-2.28	0.66	2.81
30	0.78	-4.02	-2.59	-1.10	1.48
31	1.79	-2.47	-1.10	0.60	2.02
32	2.31	-2.43	-0.99	0.27	1.91
33	0.74	-4.40	-3.12	-0.68	1.88
34	0.70	-3.37	-0.79	1.45	4.16
35	1.37	-2.88	-1.32	0.16	1.88
36	2.13	-2.52	-1.31	-0.03	1.38
37	0.71	-4.06	-2.67	0.46	2.51
38	0.85	-2.44	-0.05	2.62	4.12
39	1.87	-2.27	-0.59	1.06	2.21
40	1.42	-2.59	-1.36	0.24	1.99
41	0.72	-4.00	-2.81	-0.88	1.94
42	1.99	-2.17	-0.40	0.74	2.23
43	2.09	-2.29	-0.68	0.38	2.04
44	1.31	-2.85	-1.60	0.32	1.80
45	0.94	-2.59	-0.26	1.82	3.49
46	0.91	-3.64	-2.27	0.27	2.19
47	1.39	-2.76	-1.01	0.64	2.37
48	1.23	-2.62	-1.21	0.52	1.91

Table 3
Descriptive Statistics of Estimated Item Parameters of the 48-Item Pool, 16-Item Conventional Test and the 6-Item Routing Test

Parameter	48-Item Pool		Conventional Test		Routing Test	
	Mean	SD	Mean	SD	Mean	SD
a	1.23	0.55	1.11	0.62	2.13	0.18
b_1	-3.59	1.80	-4.09	2.12	-2.44	0.12
b_2	-1.84	1.44	-2.06	1.66	-1.21	0.36
b_3	0.13	1.24	-0.02	1.25	0.01	0.30
b_4	2.38	1.31	2.60	1.82	1.67	0.37

Table 4
Graded Response Model Descriptive Statistics for the Item Parameters of the 11 10-Item Testlets

	Attitude Trait Levels										
	-2.5	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0	2.5
Mean(a)	1.71	1.72	1.61	1.61	1.61	1.61	1.57	1.60	1.63	1.72	1.62
SD(a)	0.25	0.22	0.22	0.27	0.31	0.26	0.23	0.25	0.25	0.22	0.24
Mean(b1)	-2.46	-2.37	-2.23	-2.22	-2.22	-2.29	-2.54	-2.32	-2.23	-2.12	-2.39
SD(b1)	0.19	0.20	0.56	0.57	0.59	0.58	0.19	0.45	0.57	0.48	0.82
Mean(b2)	-1.14	-1.05	-1.06	-0.88	-0.80	-1.07	-1.20	-0.95	-1.00	-0.84	-0.98
SD(b2)	0.39	0.37	0.43	0.35	0.37	0.50	0.34	0.41	0.47	0.41	0.61
Mean(b3)	0.45	0.52	0.60	0.70	0.71	0.54	0.47	0.72	0.69	0.76	0.79
SD(b3)	0.31	0.36	0.57	0.59	0.59	0.55	0.18	0.31	0.56	0.53	0.45
Mean(b4)	2.01	2.04	2.05	2.25	2.26	2.05	2.04	2.09	2.05	2.17	2.30
SD(b4)	0.19	0.18	0.28	0.29	0.30	0.30	0.21	0.23	0.30	0.25	0.28

Table 5
Inaccuracy as a Function of Attitude Trait Level for the 16-Item Conventional Test and the Two-Stage Test

Attitude Trait Level	Inaccuracy	
	Conventional Test	2-S Test
-2.5	0.60	0.30
-2.0	0.28	0.18
-1.5	0.26	0.14
-1.0	0.31	0.18
-0.5	0.22	0.13
0.0	0.22	0.14
0.5	0.23	0.13
1.0	0.24	0.13
1.5	0.23	0.14
2.0	0.29	0.16
2.5	0.31	0.20
3.0	0.57	0.34

Table 6
Root Mean Square Difference as a Function of Attitude Trait Level for 16-Item Conventional Test and the Two-Stage Test

Attitude Trait Level	Root Mean Square Difference	
	Conventional Test	2-S Test
-2.5	0.66	0.36
-2.0	0.34	0.20
-1.5	0.31	0.16
-1.0	0.39	0.24
-0.5	0.31	0.17
0.0	0.29	0.19
0.5	0.29	0.17
1.0	0.31	0.16
1.5	0.29	0.18
2.0	0.35	0.19
2.5	0.37	0.24
3.0	0.63	0.48

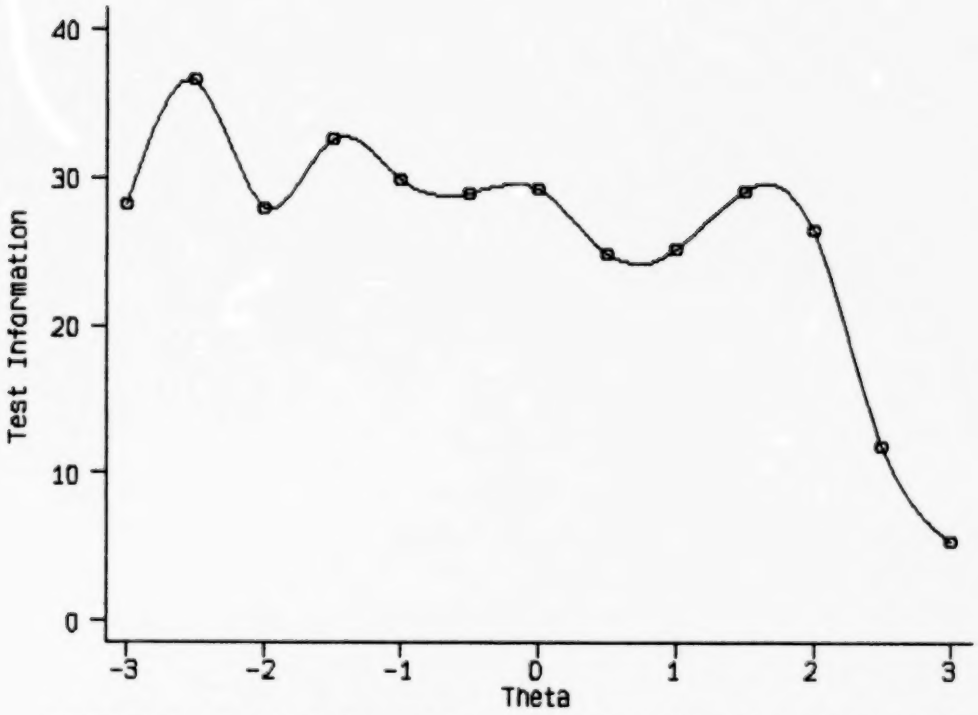

FIGURE 1 Test information function for the item pool.

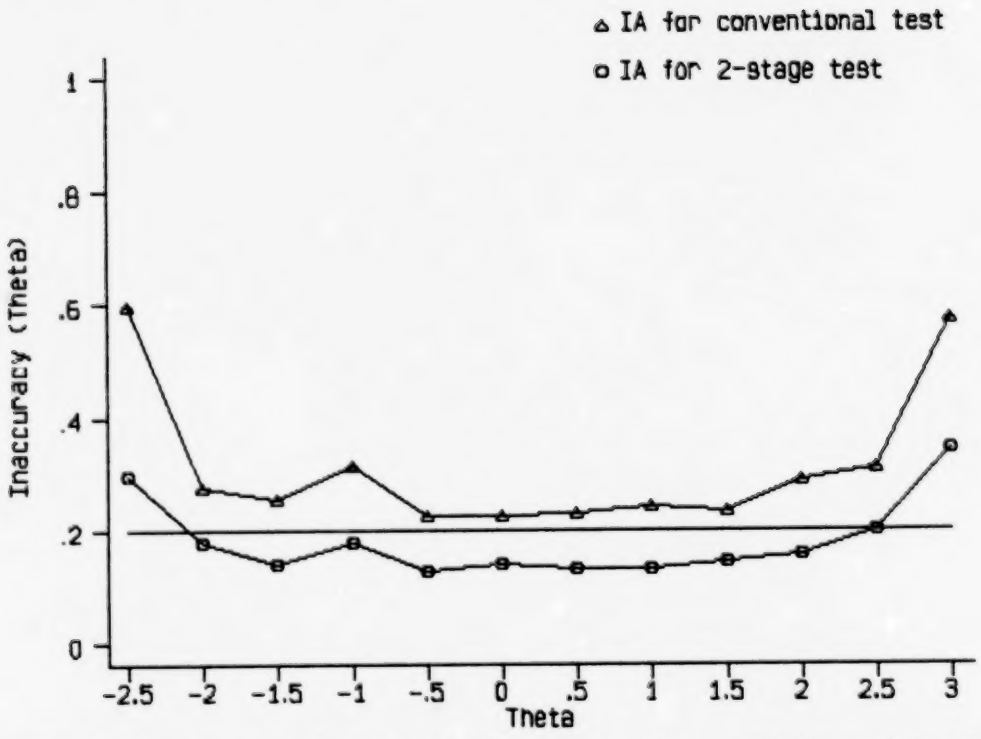

FIGURE 2 Inaccuracy for 16-item conventional and two-stage tests.

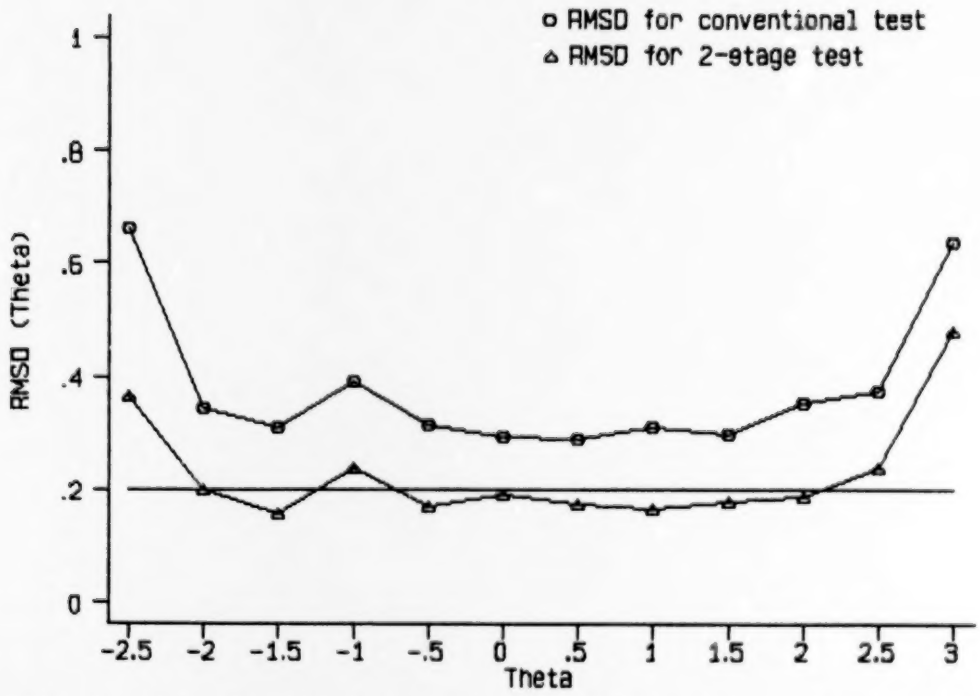

FIGURE 3 RMSD for 16-item conventional and two-stage tests.